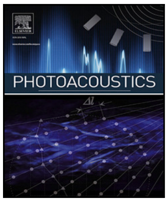
Photoacoustics 38 (2024) 100609

Contents lists available at ScienceDirect

Photoacoustics

journal homepage: www.elsevier.com/locate/pacs



Research article

Extractor-attention-predictor network for quantitative photoacoustic tomography

Zeqi Wang, Wei Tao, Hui Zhao [*]

School of Sensing Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Quantitative photoacoustic tomography
Optical inverse problem
Deep learning
Image reconstruction
Absorption coefficient estimation

ABSTRACT

Quantitative photoacoustic tomography (qPAT) holds great potential in estimating chromophore concentrations, whereas the involved optical inverse problem, aiming to recover absorption coefficient distributions from photoacoustic images, remains challenging. To address this problem, we propose an extractor-attention-predictor network architecture (EAPNet), which employs a contracting–expanding structure to capture contextual information alongside a multilayer perceptron to enhance nonlinear modeling capability. A spatial attention module is introduced to facilitate the utilization of important information. We also use a balanced loss function to prevent network parameter updates from being biased towards specific regions. Our method obtains satisfactory quantitative metrics in simulated and real-world validations. Moreover, it demonstrates superior robustness to target properties and yields reliable results for targets with small size, deep location, or relatively low absorption intensity, indicating its broader applicability. The EAPNet, compared to the conventional UNet, exhibits improved efficiency, which significantly enhances performance while maintaining similar network size and computational complexity.

1. Introduction

Photoacoustic tomography (PAT) is a rapidly advancing imaging modality with rich optical contrast and high acoustic resolution, even at depths reaching several centimeters [1–3]. The photoacoustic (PA) effect states that biological media irradiated by pulsed laser light can absorb photon energy and subsequently generate local acoustic pressure increases through thermoelastic expansion [4]. This effect is universally observed in both endogenous chromophores, such as oxygenated and deoxygenated hemoglobin, and exogenous contrast agents like nanoparticles [5–7]. From the PA signals detected by ultrasonic transducers placed around the object's surface, PA images can be reconstructed by solving a well-established acoustic inverse problem [8–10]. Quantitative photoacoustic tomography (qPAT) aims to estimate chromophore concentrations from multispectral PA images, and the key challenge is to solve the optical inverse problem referring to recovering absorption coefficients μ_a from raw PA images p_0 [11,12]. Specifically, p_0 is proportional to the product of μ_a and the local fluence Φ, while only the former is directly related to the concentration of chromophores in spectroscopic analysis [13–15].

In an earlier study, Cox et al. introduced a fixed-point iteration algorithm to recover the μ_a distribution by iteratively solving a system of equations [16]. On this basis, Zhang et al. [17] recently developed a pixel-wise reconstruction method based on a two-step iterative algorithm (TSIA), yielding noteworthy results in both simulated and experimental data. A more rigorous strategy is to employ an optimization framework, in which desired optical properties are iteratively updated, typically absorption and scattering coefficients, until the difference between the result obtained from forward modeling and measured data is minimized [18–21]. Although with high theoretical accuracy, methods within this category are computationally intensive and time-consuming, rendering real-time measurements nearly unfeasible. Some researchers attempted to directly measure fluence distributions through other techniques, such as diffuse optical tomography [22] and acousto-optic theory [23], while these methods require additional devices in a standard PA imaging system.

In the last decade, deep neural networks, especially convolutional neural networks (CNN), have made significant strides, emerging as promising tools in biomedical imaging [24–33]. In the context of qPAT, models adopt an end-to-end supervised learning manner and automatically learn the abstract feature representations from input data [34,35]. Cai et al. [29] initially extended a residual UNet to this field and gained promising results. Luke et al. [30] introduced an O-Net architecture consisting of two parallel UNets networks dedicated to the tasks of blood oxygen saturation (sO_2) estimation and vascular segmentation, respectively. Bench et al. [31] extended the O-Net framework to accommodate the 3-dimensional(3-D) nature of the PA process

[*] Corresponding author.
E-mail addresses: wangzeqi7@sjtu.edu.cn (Z. Wang), taowei@sjtu.edu.cn (W. Tao), huizhao@sjtu.edu.cn (H. Zhao).

https://doi.org/10.1016/j.pacs.2024.100609
Received 22 November 2023; Received in revised form 18 April 2024; Accepted 20 April 2024
Available online 24 April 2024

by using 3-D neural units. Li et al. [32] proposed a dual-path network comprising two UNets to estimate the absorption coefficient and fluence, respectively. Real images of fluence, absorption coefficient, and initial pressure were all used for supervised training, resulting in a joint loss function that more effectively guides parameter optimization. Zou et al. [33] integrated the anatomical features extracted from the ultrasound image by a pre-trained ResNet-18 in a standard UNet to enhance the network performance on estimating μ_a.

However, these models are founded on a UNet architecture [36]. The vanilla UNet comprises a contracting–expanding structure and a linear projection layer. Although UNet has shown impressive performance in previous studies, its nonlinear modeling capacity exclusively depends on the contracting–expanding structure. This may potentially hinder its efficiency in addressing highly nonlinear regression tasks. While increasing the width and depth of the contracting–expanding structure can compensate for this drawback, it also leads to a proliferation in network complexity. Additionally, convolutional attention mechanisms [37–39] have been widely proven to improve model performance in various imaging modalities (e.g. PAT [26,27] and computed tomography [28]). However, there is a lack of related research in the qPAT domain.

On the other hand, inspired by the pixel imbalance issue in dense object detection [40,41], we find that the mean squared error (MSE) loss [29,33], a plain per-pixel loss function commonly used in previous qPAT research, may result in the network delivering adequate performance only in specific imaging regions. Specifically, regions of high absorption or large sizes contribute the majority of the MSE loss value, exerting a dominant influence on the adjustment of network parameters. Consequently, predicted values associated with other regions are not adequately optimized during training. This optimization imbalance can cause severe spatial variations in accuracy across the output images, rendering corresponding networks unsuitable for applications involving targets of low absorption or small size. Furthermore, training with a biased focus on specific regions can result in an incomplete learning of input data, hindering the network from capturing fundamental features, and consequently deteriorating both its performance and generalization capability.

To address the above limitations, we propose a deep learning-based method utilizing an extractor-attention-predictor network architecture (EAPNet) and a balanced loss function (Bloss) for the optical inverse problem. Within the EAPNet, a contracting–expanding structure functions as an extractor for learning contextual representations containing global dependency information. Meanwhile, a pixel-wise multilayer perceptron (MLP) serves as a predictor, establishing the mapping relationship between latent features and the desired output, which can efficiently reinforce the network's nonlinear modeling capacity on a per-pixel basis. Within the E-P structure, a spatial attention module is employed to improve the capturing and utilization of important information in input images. Moreover, the Bloss function normalizes the impact of absorption intensity and size of each region on the loss value by utilizing a joint weighting factor to enhance the consistency of accuracy across the entire predicted image. Simulated and real-world validations demonstrate that our method achieves the best image-wise quantitative scores in almost all experimental scenarios and broader applicability to targets with diverse properties. Additionally, it significantly outperforms other competing methods in qualitative assessment, providing clean edges and superior agreement with ground truth images. The EAPNet proves to be more suitable and efficient compared to UNet, enhancing the accuracy of predicted μ_a images while maintaining similar network parameters and computational complexity.

2. Method

2.1. Physical fundamentals

PA images represent initial acoustic pressure distributions $p_0(\vec{r})$ generated inside media. The pressure p_0, at a given point $\vec{r}_0$, can be formulated as:

$$p_0(\vec{r}_0) = \Gamma \mu_a(\vec{r}_0)\, \Phi\left[\vec{r}_0, \mu_a(\vec{r}), \mu_s(\vec{r}), g(\vec{r})\right], \tag{1}$$

where $\mu_a(\vec{r}_0)$, $\mu_s(\vec{r}_0)$, $g(\vec{r}_0)$ and $\Phi(\vec{r}_0)$ are the local absorption coefficient, scattering coefficient, anisotropy, and fluence, respectively. Γ denotes the Grüneisen coefficient, which is typically assumed to be a constant of 1. Then, p_0 is directly equal to the product of Φ and μ_a. The fluence map $\Phi(\vec{r}_0)$ can be described by a light propagation model such as the well-established radiative transfer equation [4,42] or its approximate forms (e.g., the δ-Eddington model [43] and diffusion equation [44]). For simplicity, μ_s and g are typically treated as homogeneous constants since their spatial variations, compared to μ_a, within biological tissues are relatively small [17]. Then, Eq. (1) can be expressed in a simplified form:

$$p_0(\vec{r}_0) = \mu_a(\vec{r}_0)\, \Phi(\mu_a(\vec{r})) = F(\mu_a(\vec{r})), \tag{2}$$

where $F(\cdot)$ represents the mapping function from μ_a to p_0, describing the forward process of the PA effect. Mathematically, the optical inverse problem in qPAT aims to find an inverse operator F^{-1} that satisfies $\mu_a^{recon}(\vec{r}_0) = F^{-1}(p_0(\vec{r}))$.

The operator F^{-1} is featured with high nonlinearity, due to Φ depends on μ_a in an intricate manner [12,45]. Moreover, $p_0(\vec{r}_0)$ has a global dependency on the μ_a image, stemming from the Φ term (Eq. (2)). It suggests that $\mu_a^{recon}(\vec{r}_0)$, in turn, has a global dependency on the p_0 image. The coupling of high nonlinearity and global dependency presents a challenge in solving F^{-1}.

2.2. EAPNet architecture

The proposed extractor-attention-predictor network (EAPNet) is shown in Fig. 1 and can be formulated as:

$$\mu_a^{recon}(\vec{r}_0) = P(\vec{v}_{sa}(\vec{r}_0)) = P(A(\vec{v}(\vec{r}_0))), \tag{3}$$

$$\vec{v}(\vec{r}_0) = E(p_0(\vec{r})), \tag{4}$$

where $E(\cdot)$, $A(\cdot)$, and $P(\cdot)$ denote the operators of extraction, attention module and prediction, respectively. The extractor adopts a contracting–expanding structure, in which the contracting path learns multiscale spatial dependencies by gradually expanding the receptive field through downsampling feature maps. After aggregating the information received from skip connections, the expanding path outputs a feature map $\vec{v}(\vec{r})$ containing contextual representation vectors of all pixels. The feature map is then proceeded by a convolutional block spatial attention module (sAM) [37,38,46], to generate an attention-augmented feature map $\vec{v}_{sa}(\vec{r})$. Finally, a multilayer perceptron (MLP) is utilized to establish the mapping function between the $\vec{v}_{sa}(\vec{r})$ and μ_a on a per-pixel basis, which is mathematically a vector-to-pixel regression task.

Within the sAM, two maps (M_{mp} and M_{ap}) that aggregate channel information [37] are first extracted by conducting the max-pooling and average-pooling operations on $\vec{v}(\vec{r})$ along the channel dimension. Then, the spatial attention map M_{sa} is derived by sequentially using a 5 × 5 convolution layer $f^{5\times5}$ and a softmax activation function σ to the concatenated M_{mp} and M_{ap}, which is expressed as:

$$M_{sa} = \beta \times \sigma\left(f^{5\times5}\left[M_{mp}; M_{ap}\right]\right), \tag{5}$$

where β is a hyper-parameter. We empirically designate β as 0.5 in this study, which appears to be an optimal value as presented in the section "Parameters selection" of the supplementary materials. The attention-augmented feature map $\vec{v}_{sa}(\vec{r})$ is eventually derived by dot-multiplying the raw $\vec{v}(\vec{r})$ with the attention map M_{sa}. Compared with the seminal work in Ref. [37], there are several modifications in our attention module. Firstly, we do not employ channel attention due to the absence of discernible improvements in the associated experiment. Moreover, since our sAM operates on the feature map from the extractor, which

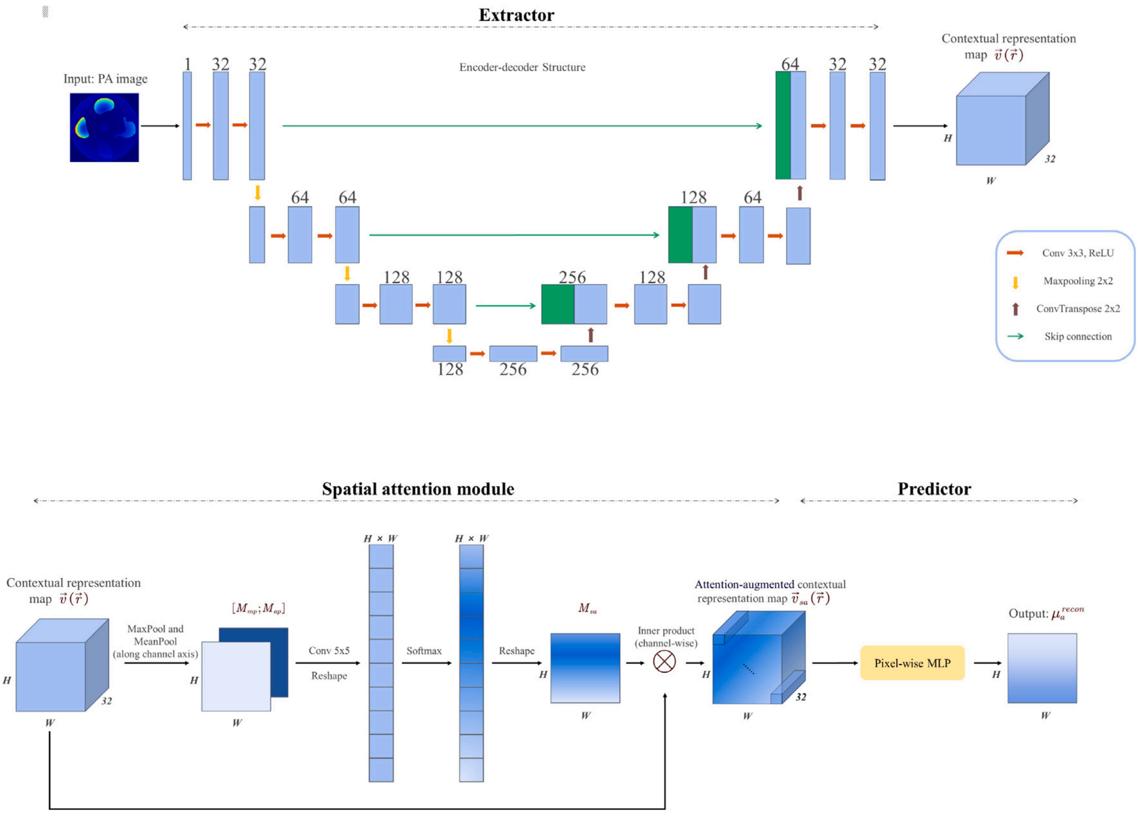


Fig. 1. Overview of the proposed EAPNet architecture, which can be divided into three successive phases: extractor, spatial attention module, and predictor. The output feature map in the first row is fed into the second row as the input. H and W denote the height and width of feature maps. The numbers annotated alongside the network indicate their corresponding number of channels. The MLP consists of two hidden layers, each having 16 channels. It maps $\vec{v}_{sa}(\vec{r})$ to $\mu_a^{recon}(\vec{r})$ as a vector-to-pixel regression task.

already integrates contextual information, the convolutional layer in our sAM utilizes a reduced kernel size of 5 × 5. Particularly, we employ a softmax function to activate our sAM because it computes spatial attention weights by considering the entire image, unlike the commonly used sigmoid function [28,37,38], which computes attention weights independently on a per-pixel basis. Specifically, the Softmax produces the output of each pixel based on its relative intensity in the input data. Since the attention map M_{sa} sums to a constant value of 1, we argue that utilizing the softmax can introduce a competitive attention mechanism. This mechanism benefits spatial information interaction, enhancing the network's capacity to distinguish and leverage informative pixels.

2.3. Loss function

To address the unbalanced optimization issue during training, as discussed in Section 1, we adopt a balanced loss function (Bloss), which, mathematically, is a form of weighted MSE loss and is expressed as:

$$\mathcal{L}_{Bloss} = \frac{1}{n}\sum_{i=1}^{n}\alpha_m\left(\frac{\mu_a^{recon} - \mu_a^{truth}}{\mu_a^{truth}}\right)^2, \tag{6}$$

where n and m denote the number of pixels and the tissue index, respectively. α_m is a factor corresponding to the mth tissue. Within the Bloss, the absolute error is divided by the μ_a^{truth} to normalize the absorption intensity. Additionally, the α_m set by the reciprocal of tissue-wise volume fractions is employed to balance the contributions of tissues with varying pixel occupancies.

2.4. Network training

All models shared the same training scheme and were implemented using PyTorch on an NVIDIA GTX 3090Ti graphics card. The ADAM optimizer was employed with an initial learning rate of 10^{-4}. All models were trained for 100 epochs with a batch size of 32. The source code is available at https://github.com/WZQ7/EAPNet.

Table 1
A brief summary of all simulation datasets.

Datasets	Major characteristics
RS	Graded absorption ranges; a deep located and small tissue.
Va	Complex vessel structures.
ST	Sparsely distributed targets.
AR	Low image quality; acoustic noise and artifacts.

3. Experiment

3.1. Simulation data generation

We generated four simulation datasets, as outlined in Table 1, to consider multiple application scenarios. Fluence distributions are simulated using MCXLAB [47]. Each simulation process lasts for one nanosecond to sufficiently capture contributions from scattered photons, and a total of 10^8 photons are emitted. Photons leaving the domain are terminated. The simulation domain of the vasculature (Va) and sparse target (ST) datasets is 128 × 64 pixels with a single line source placed along the surface. The random shape (RS) and acoustic reconstruction (AR) datasets are simulated within a 256 × 256 pixels field, and line sources are placed along all four sides to ensure wide-field illumination. For all datasets, the pixel length is 0.1 mm, and the Grüneisen parameter is assumed constant at 1, leading to p_0 being equal to the product of μ_a and Φ. Each dataset comprises 4000 annotated image pairs, divided into training, validation, and testing sets in an 8:1:1 ratio. Networks are trained and evaluated on their respective

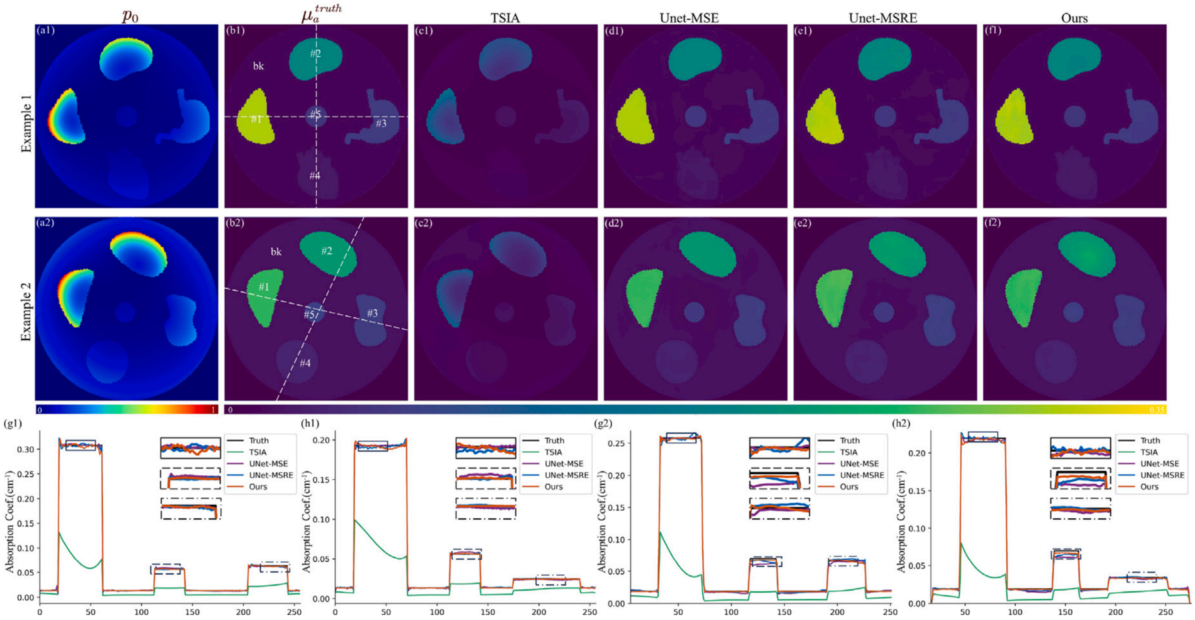


Fig. 2. Example images of the RS dataset. Tag numbers indicate the example index. (a) Raw PA images. (b) Ground truth, μ_a^{truth}. (c–f) μ_a^{recon} obtained from different methods. (g) Profiles along the horizontal white dotted line in (b). (h) Profiles along the vertical white dotted line in (b). In (g) and (h), the contained insets show the enlarged parts within the black boxes that are labeled by corresponding line styles. ''bk'' is the abbreviation of ''background''. Quantitative metrics of example images are detailed in Table S4.

datasets. The validation set is utilized for selecting hyperparameters. Before being fed into a network, all input images were normalized to their maximum values.

(i) Random shape dataset. The dataset consists of circular phantoms with a physical radius of 25.6 mm. Each phantom comprises four random shapes mimicking biological tissues and a small circle. Typical geometries are illustrated in Fig. 2. Tissue positions are determined based on prescribed locations with random perturbations, to enrich the data distribution. Different absorption ranges are assigned to these tissues, as detailed in Table S3 of the supplementary materials. Throughout the entire phantom, a reduced scattering coefficient of 1 mm^{-1} and an anisotropy parameter of 0.9 are consistently utilized.

(ii) Vasculature dataset. The purpose of this dataset is to assess our method's performance in retrieving intricate structures and subtle textures. Each phantom is constructed based on a two-layer skin model measuring 128 × 64 pixels. The top three rows of pixels represent the epidermis layer, while the remainder constitutes the dermis layer. The vessels utilized are sourced from the publicly available FIVES dataset comprising 800 fundus photographs [48]. Each raw image is downsampled and then evenly divided into nine patches of 128 × 64 pixels, with a 25% overlap between adjacent patches for data augmentation. After a manual review to remove those with insufficient or unclear vessels, each remained patch is implanted in an individual dermis layer. All patches are used only once, ensuring a distinct vascular structure for each phantom. Optical properties are determined using empirical equations and parameters from Refs. [31,49] to enhance data realism. Additionally, pixel-wise variations in absorption are introduced to augment textural richness.

(iii) Sparse target dataset. This dataset aims to assess the effectiveness of our method in reconstructing sparsely distributed and small-size targets, in which the phantoms employ the identical two-layer skin model and optical parameter assignment approach as those in the Va dataset. The sparsely distributed vascular structures are extracted from a publicly available lung image dataset.[1] As shown in Fig. 4(b), the imaging field is largely occupied by the skin tissue.

(iv) Acoustic reconstruction dataset. The PA images in the three mentioned datasets represent ideal initial pressure distributions. However, real images are susceptible to corruption from factors such as system noise and imperfect detection conditions during the acquisition process. To account for these influences, we created the AR dataset by simulating the acoustic measurement process on the RS dataset using the MATLAB toolbox k-Wave [50]. The p_0^{recon} images are reconstructed from the simulated time-domain PA signals using a time reversal algorithm [9]. Implementation details are available in the supplementary materials.

3.2. Phantom experiment

To verify the feasibility of our method in real-world scenarios, we applied it to phantom images acquired by a commercial photoacoustic tomography system mentioned in Section 3.4. The geometries of real phantoms are close to that of the RS dataset. Each phantom, with a radius of 9.6 mm, contains five inclusions, as shown in Fig. 6. Based on the recipe in Ref. [51], we first mixed 1% agar powder solution (A-1296, Sigma) with 20% Intralipid to produce a base solution. The solutions constituting different tissues were prepared by adding corresponding concentrations of India ink into the base solution. We provide more details on preparing phantoms in supplementary materials.

In the acquisition process, we initially averaged the raw time-domain signals from ten repetitive measurements to improve the signal-to-noise ratio (SNR). Subsequently, a high-accuracy model-based iterative reconstruction algorithm [8,52] processed the averaged signals and generated raw images. Finally, these images were normalized to their maximum values to remove scalar factors associated with system parameters.

All networks used in this phantom experiment were trained on the AR dataset. For quantitative assessment of network performance, we annotated several phantom images with μ_a determined by the following experimental procedure. First, each used material was poured into a mold consisting of two glass slides and a U-shaped spacer to prepare a corresponding slab-shaped test sample. Subsequently, a spectrophotometer (LAMBDA 950, PerkinElmer) equipped with an integrating sphere was used to measure the total transmittance and reflectance of these samples. Optical properties were then derived via the inverse adding-doubling algorithm [53], assuming a constant anisotropy of 0.9.

[1] Public Lung Image Database, http://www.via.cornell.edu/lungdb.html.

Table 2
Quantitative evaluation (MEAN ± SD) for 400 test images from the RS dataset. SD denotes the standard deviation of a single metric across the test set. As explained in Section 3.6, the tissue-wise metrics (i.e., $\mathrm{MRE_m}$ and $\mathrm{PSNR_m}$) refer to the result within a specific tissue region. The tSD refers to the standard deviation across the mean values of tissue-wise metrics of a given method. The tMRE and tPSNR are used as image-wise metrics. The regions corresponding to the labels “bk” and “#1”–“#5” can be found in Fig. 2.

	Algorithms	Tissue-wise metrics						tSD	tMRE (top) tPSNR (bottom)
		bk	#1	#2	#3	#4	#5		
|Relative error|	Unet-MSE	0.058 ± 0.013	0.007 ± 0.003	0.009 ± 0.003	0.031 ± 0.009	0.054 ± 0.026	0.064 ± 0.036	0.023	0.037 ± 0.009
	Unet-MSRE	0.034 ± 0.007	0.022 ± 0.006	0.023 ± 0.009	0.026 ± 0.012	0.033 ± 0.015	0.052 ± 0.029	0.010	0.032 ± 0.007
	Ours	0.014 ± 0.006	0.016 ± 0.005	0.016 ± 0.005	0.017 ± 0.006	0.018 ± 0.007	0.021 ± 0.012	0.002	0.017 ± 0.005
PSNR (dB)	Unet-MSE	22.437 ± 1.731	40.498 ± 2.513	38.268 ± 2.419	28.673 ± 2.106	24.571 ± 3.148	24.338 ± 4.507	7.057	29.797 ± 1.428
	Unet-MSRE	27.284 ± 1.654	31.009 ± 1.995	31.230 ± 2.344	30.097 ± 2.587	28.567 ± 2.917	25.749 ± 4.064	2.010	28.989 ± 1.318
	Ours	35.517 ± 2.528	34.005 ± 2.273	33.891 ± 2.137	33.702 ± 2.255	33.241 ± 2.595	33.611 ± 3.999	0.722	33.994 ± 1.703

3.3. In vivo experiment

We conducted a preliminary validation in an in vivo scenario. A healthy nude female mouse weighing approximately 15 grams was imaged after anesthesia. Cross-sectional PA images were acquired at the abdominal region, mainly treating kidneys as the regions of interest. The acquisition adhered to the procedure outlined in Section 3.2. Animal procedures adhered to the Guidelines for Care and Use of Laboratory Animals of Shanghai Jiao Tong University, and experiments received approval from its Animal Ethics Committee. Since constructing a specialized training dataset for in vivo data is non-trivial and usually regarded as an independent research task, the networks for this initial experiment are trained using synthetic data. We explain this dataset in the supplementary materials.

3.4. PAT imaging system

All experimental data was acquired by a commercial small-animal multispectral optoacoustic tomography system (MSOT inVision256, iThera Medical). The light source consists of a pulsed Nd:YAG laser (9 ns pulse width and 10 Hz repetition rate) and an optical parametric oscillator, enabling a multi-wavelength irradiation ranging from 680 to 1200 nm. Incident light is delivered to sample surfaces via a ten-arm fiber bundle to achieve 360-degree illumination in the imaging plane. PA waves are recorded by a ring-shaped transducer array consisting of 256 cylindrical-focused elements characterized by a central frequency of 5 MHz and a bandwidth of 60%. These transducer elements are arranged in a 270-degree arc with a radius of 40.5 mm.

3.5. Comparative methods

For comparison, we employed the two-step iterative algorithm (TSIA) [17] as a conventional method and UNet (Fig.S1) as the network benchmark. UNet was trained with mean squared error (UNet-MSE) and mean squared relative error loss (UNet-MSRE) to explore the impact of different loss functions. We provide details on the implementation of comparative methods in supplementary materials. Notably, the comparison experiments focus on the superiority of EAPNet as a fundamental architecture, and only UNet was employed here because, in the context of qPAT, most models are built upon UNet. Strategies, including employing multiple UNets [30,32] and using modified convolution units (such as the residual block [29] and 3-D block [31]), could be readily applied to EAPNet and achieve corresponding enhancements from an improved performance baseline, while these are beyond the scope of our investigation.

3.6. Evaluation metrics

Evaluation metrics in previous studies may inadequately reflect the overall accuracy of predicted images in some cases. Per-pixel metrics, such as mean absolute error and mean relative error (MRE) [32,54], are inclined to reflect the error levels within those large tissues. The peak signal-to-noise ratio (PSNR), directly computed from the maximum μ_a, may result in a biased value when there is a significant disparity in μ_a between different tissues [55]. To address these limitations, two metrics adopted in this study, tMRE and tPSNR, are computed on a per-tissue basis. Specifically, tissue-wise metrics ($\mathrm{MRE_m}$ or $\mathrm{PSNR_m}$) are computed initially, where m represents the tissue index. Evaluation metrics for the predicted image (tMRE or tPSNR) are subsequently derived by averaging these tissue-wise metrics. tMRE is formulated as:

$$\mathrm{MRE_m} = \frac{1}{N_m}\sum_{\vec{r}\in\Omega_m}\frac{\left|\mu_a^{recon}-\mu_a^{truth}\right|}{\mu_a^{truth}}, \tag{7}$$

$$\mathrm{tMRE} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{MRE_m}, \tag{8}$$

where Ω_m and N_m denote the domain of the mth tissue and the number of pixels within it, respectively. M is the number of tissue types. tPSNR can be formulated as:

$$\mathrm{MSE_m} = \frac{1}{N_m}\sum_{\vec{r}\in\Omega_m}\left(\mu_a^{recon}-\mu_a^{truth}\right)^2, \tag{9}$$

$$\mathrm{PSNR_m} = 20\log 10\left(\frac{\max\left(\mu_{a,m}\right)}{\sqrt{\mathrm{MSE_m}}}\right), \tag{10}$$

$$\mathrm{tPSNR} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{PSNR_m}, \tag{11}$$

where $\max\left(\mu_{a,m}\right)$ indicates the maximum possible value within Ω_m.

Moreover, the standard deviation across tissue-wise metrics (the mean values of $\mathrm{MRE_m}$ or $\mathrm{PSNR_m}$ on the test set are adopted here) is employed to assess the consistency of prediction accuracy among various tissues for each method, and we denote it as tSD.

In addition, we introduce a constrained generalized contrast-to-noise ratio (gCNR) [55–57], denoted as CgCNR, to evaluate the efficacy of our method in recovering intricate or fine structures. CgCNR is exclusively applied to the ST and Va datasets since the RS and AR datasets only consist of simple structures.

Previous studies have proven that gCNR is directly associated with the separability of the target from the background, offering a better metric of distinction (or sharpness) compared to contrast and contrast-to-noise ratio [57]. However, gCNR fails to consider the agreement between the target distribution and the ground truth, and only a high distinction does not match the goal of qPAT. CgCNR is a modification of the standard gCNR and can be expressed as follows:

$$\mathrm{CgCNR} = \min_{\varepsilon}\left\{1-\left(\int_{-\infty}^{\varepsilon}p_i(x)\,dx+\int_{\varepsilon}^{\infty}p_o(x)\,dx\right)\right\}, \quad \text{s.t. } \varepsilon > \varepsilon_L, \tag{12}$$

where x denotes the pixel value. $p_i(x)$ and $p_o(x)$ represent the probability density functions of the target and background, specifically referring to the vascular and dermal areas in this context. From Eq. (12), CgCNR shares an identical optimization problem with gCNR but poses a lower limit ε_L for ε as an additional constraint. Under this constraint, the CgCNR only evaluates the separability of target pixels sufficiently close to the ground truth. Target pixels whose predicted values below ε_L are

Table 3
Quantitative evaluation (MEAN ± SD) for 400 test images from the Va dataset. A detailed explanation of all used metrics can be found in Table 2.

	Algorithms	Tissue metrics			tSD	tMRE (top) tPSNR (bottom)
		Epidermis	Dermis	Vessel		
|Relative error|	Unet-MSE	0.023 ± 0.028	0.349 ± 0.150	0.186 ± 0.055	0.133	0.186 ± 0.058
	Unet-MSRE	0.024 ± 0.023	0.065 ± 0.010	0.254 ± 0.155	0.100	0.114 ± 0.051
	Ours	0.062 ± 0.030	0.070 ± 0.013	0.081 ± 0.030	0.008	0.071 ± 0.020
PSNR (dB)	Unet-MSE	36.142 ± 8.917	4.160 ± 3.522	13.832 ± 2.271	13.392	18.045 ± 2.812
	Unet-MSRE	31.777 ± 4.944	22.051 ± 1.373	11.462 ± 4.168	8.296	21.763 ± 1.540
	Ours	22.700 ± 3.525	20.468 ± 2.929	20.168 ± 3.169	1.130	21.112 ± 2.738

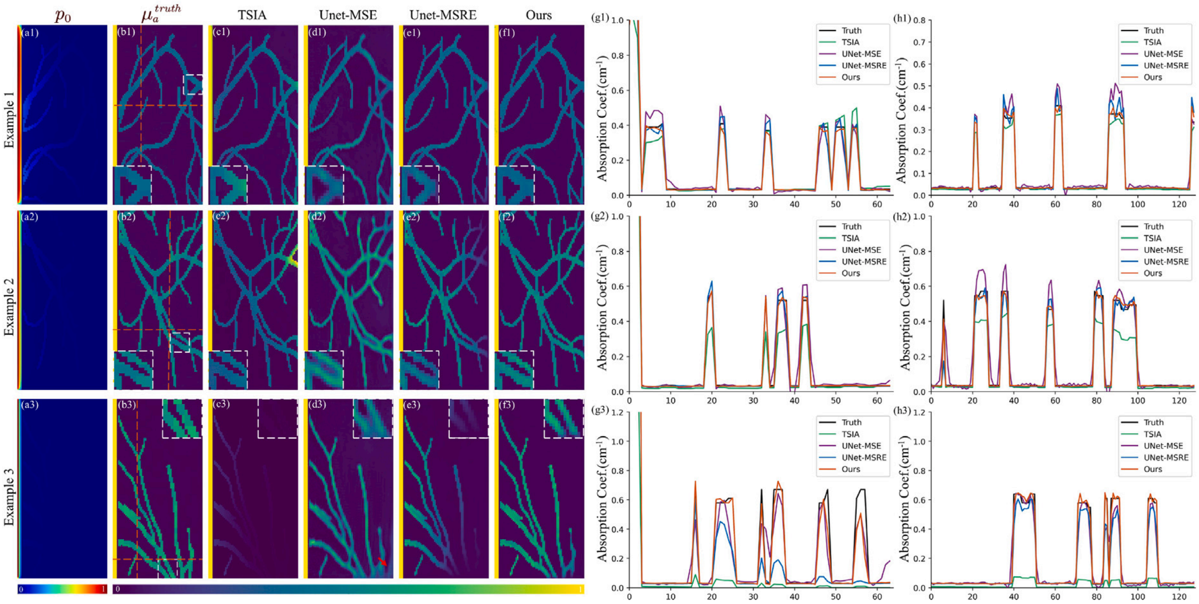


Fig. 3. Example images of the Va dataset. Tag numbers indicate the example index. (a) Raw PA images. (b) Ground truth, μ_a^{truth}. (c–f) μ_a^{recon} obtained from different methods. The ROI of each example is delineated by a white dotted box in (b) and magnified for better visualization. These ROIs are also used for calculating CgCNR. (g) Profiles along the horizontal red dotted line in (b). (h) Profiles along the vertical red dotted line in (b). Quantitative metrics of example images are detailed in Table S4.

directly treated as false negatives because they lack practical significance for quantification applications. In our study, ε_L was set to 75% of the mean value of μ_a in vascular regions. Implementation details are available in the supplementary materials.

4. Result

4.1. Simulation data analysis

Upon a preliminary observation of Figs. 2 to 5, we can easily find that there is a depth-dependent decline in the intensities of p_0 images due to fluence attenuation. In addition, the TSIA produces acceptable results only in certain cases [Figs. 3(c1,c2) and 4(c1,c2)] because the δ-Eddington model used performs poorly in regions near light sources or with high absorption. Therefore, TSIA demonstrates inadequate robustness against variations in illumination conditions and the optical properties of the imaging medium. Due to the significantly lower accuracy of TSIA compared to DL-based methods across all datasets, we focus our comparison on DL-based methods in the following parts.

(i) Random shape dataset. As shown in Table 2, our method outperforms other algorithms significantly in terms of both tMRE and tPSNR comparisons. Moreover, tissue-wise evaluation reveals a distinct variation in the accuracy of both UNet-MSE and UNet-MSRE across different tissues. UNet-MSE's performance deteriorates as tissue absorption decreases, leading to relatively large errors in regions with low absorption (e.g., the background and tissues #3–#5), while UNet-MSRE exhibits relatively subpar performance in the small-sized tissue #5. The non-uniform accuracy causes unreliable results in practical applications, as the target tissues are not always of high absorption or large size. Our method is insensitive towards tissue properties and enables greater uniformity in accuracy across tissues, as evidenced by the minimal tSD. This also implies that our method has broader applicability.

There is no marked visual distinction among the μ_a^{recon} images of different methods (Fig. 2). The profiles and their magnified regions display subtle differences, indicating that results predicted by our method provide closer concordance with the ground truth.

(ii) Vasculature dataset. Due to the highly unbalanced contributions of different tissues to the loss value, the performance of both UNet-MSE and UNet-MSRE varies significantly across them, as shown in Table 3. While excelling in the epidermal and dermal regions, respectively, the two methods underperform in vascular regions of significant medical interest. Our method provides distinct superiority in vascular regions and also better overall accuracy, achieving the lowest tMRE, the second-highest tPSNR, and the minimal tSD.

Fig. 3 showcases three skin models with progressively enhanced epidermal absorption, representing Caucasian, Asian, and African skin types. The enlarged images display regions of interest (ROIs) highlighted by white dotted boxes in Fig. 3(b). UNet-MSE performs poorly in preserving delicate features and structures. It outputs vessels with blurred edges and may produce non-existent vessels, as indicated by the red arrow in Fig. 3(d3). In cases of relatively strong epidermal absorption, UNet-MSRE fails to adequately compensate for the effects of fluence attenuation, notably underestimating μ_a of deep vessels, thereby rendering them unobservable in the images. Our method reconstructs vascular structures with exceptional fidelity and well-defined

Table 4
Quantitative evaluation (MEAN ± SD) for 400 test images from the ST dataset. A detailed explanation of all used metrics can be found in Table 2.

	Algorithms	Tissue-wise metrics			tSD	tMRE (top) tPSNR (bottom)
		Epidermis	Dermis	Vessel		
|Relative error|	Unet-MSE	0.010 ± 0.019	0.171 ± 0.050	0.173 ± 0.058	0.076	0.118 ± 0.032
	Unet-MSRE	0.017 ± 0.014	0.045 ± 0.011	0.645 ± 0.187	0.290	0.236 ± 0.060
	Ours	0.049 ± 0.030	0.055 ± 0.016	0.076 ± 0.030	0.012	0.060 ± 0.020
PSNR (dB)	Unet-MSE	43.686 ± 9.127	13.136 ± 3.048	15.888 ± 3.032	13.799	24.237 ± 2.378
	Unet-MSRE	34.397 ± 4.380	25.700 ± 2.257	5.587 ± 2.565	12.066	21.895 ± 1.391
	Ours	24.204 ± 3.025	24.123 ± 2.551	22.444 ± 3.533	0.811	23.590 ± 2.568

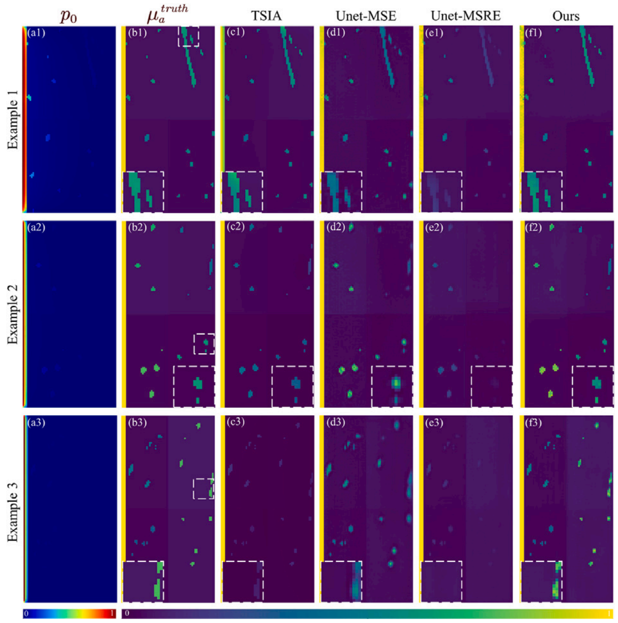


Fig. 4. Example images of the ST dataset. Tag numbers indicate the example index. (a) Raw PA images. (b) Ground truth, μ_a^{truth}. (c–f) μ_a^{recon} obtained from different methods. The ROI of each example is delineated by a white dotted box in (b) and magnified for better visualization. These ROIs are also used for calculating CgCNR. Quantitative metrics of example images are detailed in Table S4.

edges, demonstrated by the highest CgCNR score in Table S5 of the supplementary materials. The line profiles illustrate that the predicted μ_a of our method exhibits a greater degree of alignment with the ground truth. Meanwhile, UNet-MSE and UNet-MSRE are sensitive to the epidermal absorption intensity, resulting in a decrease in image quality with its escalation. In contrast, our method shows sufficient robustness to this variation.

(iii) Sparse target dataset. As illustrated in Table 4, the sparse distribution of blood vessels significantly impacts the performance of UNet-MSRE, resulting in considerable estimation biases in vascular regions. This is because MSRE only corrects for the unbalanced impact of absorption intensity on loss values. In contrast, our method utilizes the weight factor α_m to effectively address the training imbalance resulting from the difference in pixel occupancy among tissues, making it robust to sparse targets, as evidenced by significantly better performance in vascular regions and the lowest tSD.

Representative images of three skin types (Caucasian, Asian, and African) are displayed in Fig. 4. Similar to the findings in Fig. 3, UNet-MSE overly smooths vessel boundaries, sacrificing structural details. The μ_a^{recon} images produced by UNet-MSRE fail to adequately distinguish deep vessels. Our method recovers vessels of superior sharpness, which is also demonstrated by the highest CgCNR in Table S6 of the supplementary materials.

(iv) Acoustic reconstruction dataset. From Fig. 5, it is evident that the p_0^{recon} images are distorted compared to the ideal p_0 images due to noise and artifacts. Metrics for 400 test images are listed in Table 5. As expected, utilizing non-ideal input data results in reduced performance across all methods compared to the results presented in Table 2. The tMRE and tPSNR demonstrate that the overall performance of our method ranks at the top. Meanwhile, our method surpasses other methods in tSD, suggesting our method effectively provides consistent accuracy for different tissues. The profile images further confirm the accuracy advantage of our method, particularly in tissue #5 and the background.

4.2. Phantom data analysis

Fig. 6 shows the phantom experiment results. The acquired p_0 images contain severe artifacts and have weak visibility in the deep area. There is a noticeable degradation across all evaluation indicators compared to the simulation results since the networks were only trained on the AR dataset. Although possessing similar internal geometries and optical properties, the AR dataset still fails to sufficiently represent the phantom data, leading to networks lacking an understanding of the features specific to the experimental domain. Here are several specific manifestations. In our imaging system, the ring-like transducer array provides only 270-degree angular coverage around the object, leaving the top 90 degrees uncovered to accommodate the mouse holder. The limited view results in signal loss and significant reconstruction errors in neighboring regions, notably for tissue #2 and the upper regions of tissues #1 and #3. Moreover, significant errors can be observed at phantom boundaries due to mismatches in illumination conditions. Despite these challenges, it seems that all methods can output reasonable μ_a^{recon} images, suggesting that networks trained on synthetic datasets, to some extent, can be applicable to real data with sufficient similarity.

Quantitative indicators of the phantom results are shown in Table 6. Our method obtains the best scores for the tMRE and tPSNR. For the μ_a^{recon} images, our method effectively corrects out the spatial varying fluence, and tissues obtain enhanced signal homogeneity. This advantage is further shown by the profile images. Particularly for the background area with low fluence, both UNet-MSE and UNet-MSRE are inclined to underestimate μ_a, whereas our method consistently aligns closely with the reference value.

Notably, while experimentally obtained μ_a of each tissue offers a reasonable reference, uncertainties introduced by human and equipment factors are unavoidable.

4.3. In vivo data analysis

Fig. 7 shows the results of the cross-sectional abdominal images at a wavelength of 760 nm. In the acquired p_0 images, the signal is stronger on the body surface, and the visibility diminishes in the central region. Our method efficiently corrects internal fluence attenuation, restoring the arterial signal to a relatively normal magnitude. In the p_0 profiles, signals of identical organs exhibit a depth-dependent decrease, whereas they become more homogeneous in predicted images obtained from our method, as indicated by the corresponding black dotted lines and red arrows [Fig. 7(c–f)]. To enable a convenient comparison between two measurements of the same slice, we horizontally flipped example 1 (p_0

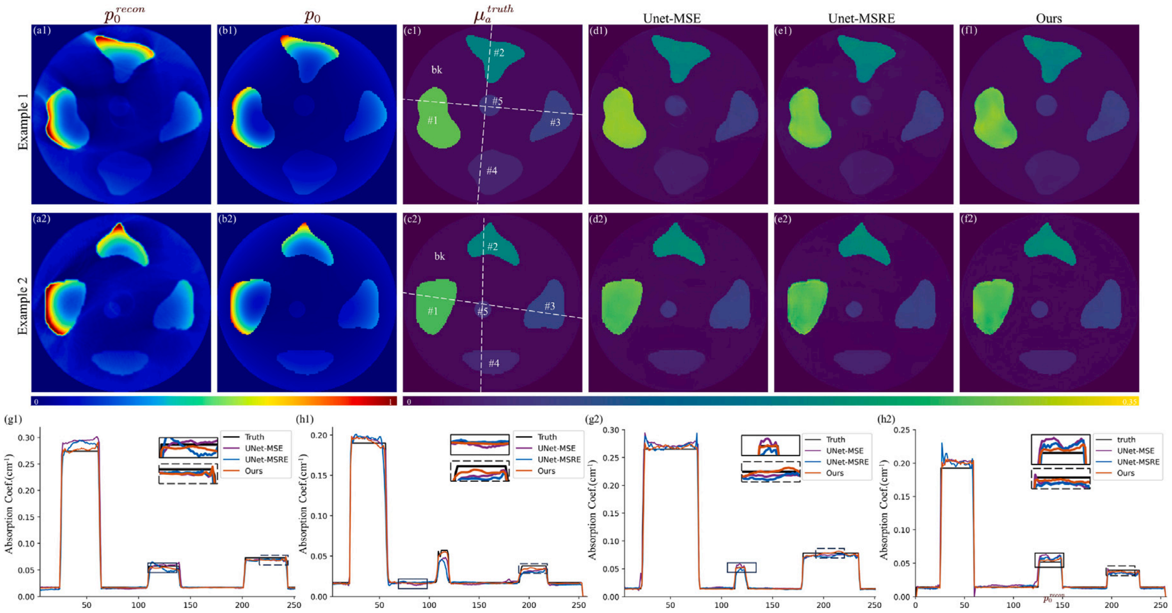


Fig. 5. Example images of the AR dataset. Tag numbers indicate the example index. (a) Reconstructed PA images as inputs, p_0^{recon}. (b) Initial pressure distributions (perfect PA images). (c) Ground truth, μ_a^{truth}. (d–f) μ_a^{recon} images obtained from different methods. (g) Profiles along the horizontal white dotted line in (c). (h) Profiles along the vertical white dotted line in (c). In (g) and (h), the contained insets show the enlarged parts within the black boxes that are labeled by corresponding line styles. "bk" is the abbreviation of "background". Quantitative metrics of example images are detailed in Table S4.

Table 5
Quantitative evaluation (MEAN ± SD) for 400 test images from the AR dataset. A detailed explanation of all used metrics can be found in Table 2. The regions corresponding to the labels "bk" and "#1"–"#5" can be found in Fig. 5.

	Algorithms	Tissue-wise metrics						tSD	tMRE (top) tPSNR (bottom)
		bk	#1	#2	#3	#4	#5		
|Relative error|	Unet-MSE	0.148 ± 0.064	0.040 ± 0.022	0.043 ± 0.025	0.092 ± 0.051	0.108 ± 0.059	0.101 ± 0.047	0.038	0.088 ± 0.022
	Unet-MSRE	0.085 ± 0.025	0.075 ± 0.025	0.071 ± 0.024	0.100 ± 0.044	0.110 ± 0.054	0.139 ± 0.050	0.023	0.097 ± 0.019
	Ours	0.035 ± 0.018	0.068 ± 0.031	0.066 ± 0.030	0.082 ± 0.042	0.072 ± 0.040	0.073 ± 0.042	0.015	0.066 ± 0.018
PSNR (dB)	Unet-MSE	12.214 ± 2.704	26.052 ± 3.277	25.764 ± 3.420	20.187 ± 3.597	18.748 ± 2.844	18.150 ± 2.844	4.749	20.186 ± 1.555
	Unet-MSRE	19.345 ± 2.242	16.678 ± 1.481	17.533 ± 1.501	17.816 ± 2.267	18.143 ± 2.952	13.991 ± 1.654	1.660	17.251 ± 1.137
	Ours	23.868 ± 2.445	19.518 ± 2.297	20.045 ± 2.241	20.243 ± 3.037	21.864 ± 3.402	21.133 ± 3.438	1.449	21.112 ± 1.592

Table 6
Quantitative evaluation of the phantom results in Fig. 6 (MEAN ± SD). A detailed explanation of all used metrics can be found in Table 2. The regions corresponding to the labels "bk" and "#1"–"#5" can be found in Fig. 6.

	Algorithms	Tissue-wise metrics						tMRE (top) tPSNR (bottom)
		bk	#1	#2	#3	#4	#5	
|Relative error|	UNet-MSE	0.253 ± 0.029	0.407 ± 0.121	0.222 ± 0.023	0.243 ± 0.075	0.137 ± 0.006	0.315 ± 0.090	0.263 ± 0.008
	UNet-MSRE	0.270 ± 0.014	0.319 ± 0.088	0.277 ± 0.026	0.284 ± 0.062	0.218 ± 0.015	0.330 ± 0.118	0.283 ± 0.015
	Ours	0.194 ± 0.073	0.165 ± 0.004	0.097 ± 0.001	0.240 ± 0.082	0.137 ± 0.002	0.110 ± 0.043	0.157 ± 0.009
PSNR (dB)	UNet-MSE	4.285 ±0.640	6.673 ± 2.174	11.333 ± 0.530	11.309 ± 1.751	15.883 ± 0.322	9.231 ± 2.216	9.786 ± 0.334
	UNet-MSRE	6.030 ± 1.307	8.833 ± 2.316	10.021 ± 0.556	9.828 ± 1.306	12.420 ± 0.598	8.286 ± 2.169	9.236 ± 0.032
	Ours	3.649 ± 1.309	13.274 ± 0.745	15.476 ± 0.321	10.958 ± 1.769	15.682 ± 0.011	14.544 ± 1.044	12.264 ± 0.071

image) to generate example 2, which, for the network, is equivalent to distinct imaging data [58,59]. The predicted values for corresponding locations from the two images show satisfactory agreement, further confirming the network's effectiveness.

It is important to mention that, as our network was not trained on specialized in vivo data, its effectiveness shown in this validation is relatively limited, and only qualitative analysis was conducted. The lack of labeled in vivo data undoubtedly hinders our network from reaching its full performance potential. We elaborate on these challenges associated with in vivo applications in Section 5.

4.4. Ablation study

An ablation study was conducted on the RS and Va datasets, examining the network architecture, loss function, and attention module:

(1) Network ablation: UNet-Bloss;

(2) Loss ablation: EAPNet with the MSE loss function (EAPNet-MSE) and EAPNet with the MSRE loss function (EAPNet-MSRE).

(3) Attention module ablation: removing the proposed sAM from EAPNet (EPNet-Bloss), and replacing our activation function with the sigmoid function in the sAM [EAPNet-Bloss (sigmoid)].

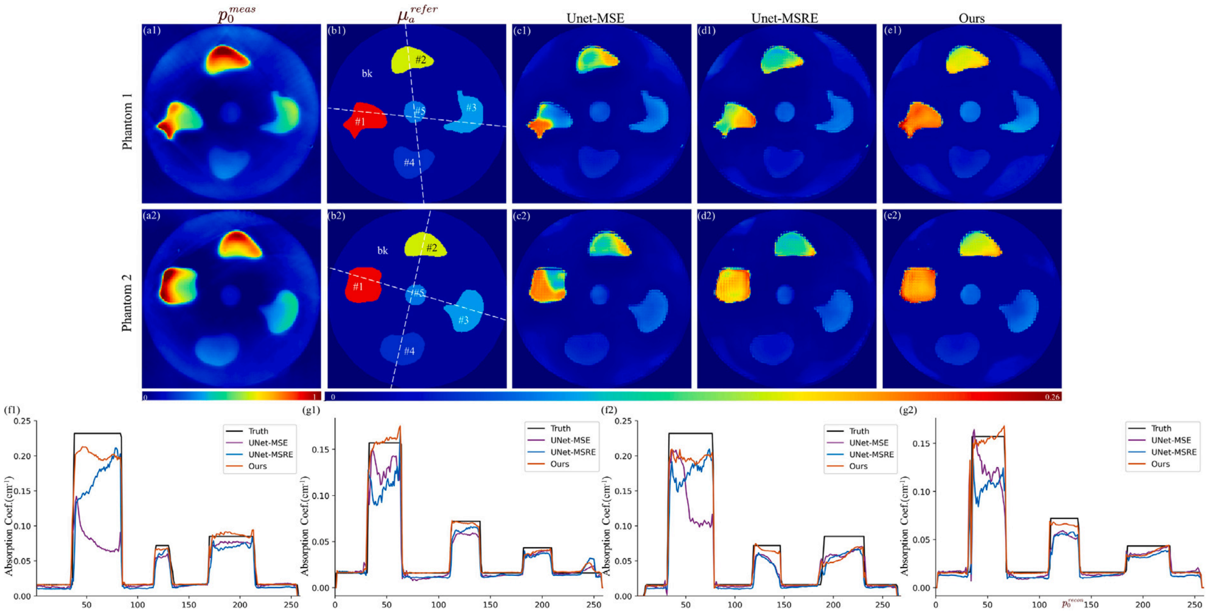


Fig. 6. Results of the phantom experiment. Tag numbers indicate the example index. (a) Experimentally acquired PA images. (b) Reference value of μ_a determined experimentally. (c–e) μ_a^{recon} images obtained from different methods. (f) Profiles along the horizontal white dotted line in (b). (g) Profiles along the vertical white dotted line in (b). "bk" is the abbreviation of "background".

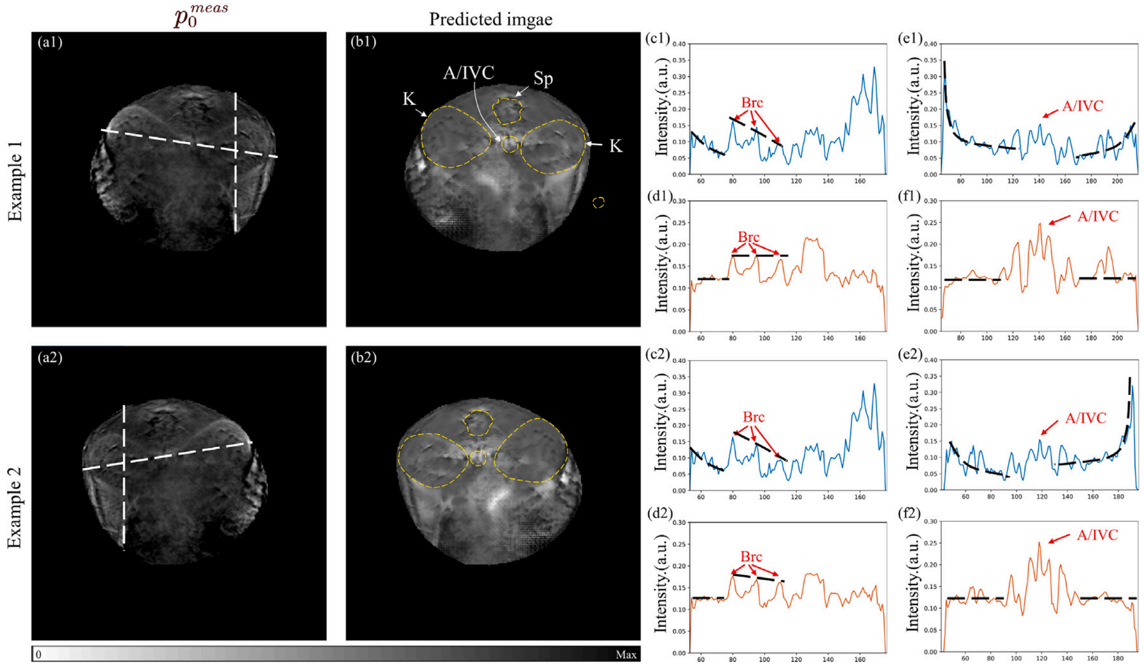


Fig. 7. Results of the in vivo experiment. Pictures with the same label number correspond to the results of one example. (a) Experimentally acquired PA images. (b) The predicted images obtained from our method. (c, e) Profiles of p_0^{meas} along the two white dotted lines in (a). (d, f) Profiles of the predicted images along the two white dotted lines in (a). Description of markers: A: artery; IVC: inferior vena cava; K: kidney; Sp: spine; Brc: boundary region of the renal cortex.

The results for the RS and Va datasets are listed in Table 7 and Table S7, respectively, which consistently demonstrate our method obtains the best evaluation metrics. This ablation study confirms that each proposed module in our method is valuable and contributes to better performance.

4.5. Network design analysis

Table S8 of the supplementary materials presents information regarding the number of parameters, floating point operations (FLOPs), and computational time for each network. Compared to the UNet, EAPNet shares a comparable network size and computational complexity, yet consistently obtains superior performance conditioning on the identical loss function, as shown by the ablation experiments.

To further demonstrate the superiority of our network architecture, we trained two more complex UNets with the Bloss on the RS dataset, one wider and one deeper. Compared to the base UNet, the channel dimension of each convolution layer in the wider UNet is doubled, and the deeper UNet includes an extra contracting block and an extra expanding block The comparison results are listed in Table S8. While

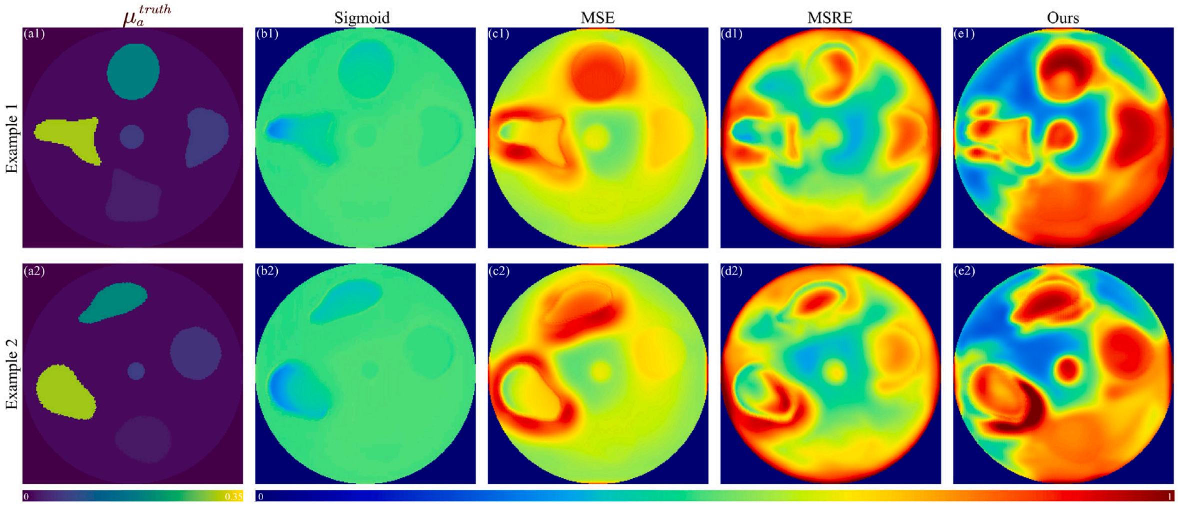


Fig. 8. Attention map analysis of two examples in the RS dataset. Tag numbers indicate the example index. (a) Ground truth, μ_a^{truth}. (b–e) Attention maps obtained from different methods, from left to right: proposed sAM with Bloss but using the sigmoid function as the activation function, proposed sAM with MSE, proposed sAM with MSRE, proposed sAM with Bloss (ours).

Table 7
Results of the ablation study on the RS dataset (MEAN ± SD). A detailed explanation of all used metrics can be found in Table 2.

	Algorithms	tMRE	tPSNR
Full method	Ours	**0.017** ± **0.005**	**33.994** ± **1.703**
Network ablation	UNet-Bloss	0.030 ± 0.006	29.567 ± 1.270
Loss ablation	EAPNet-MSE	0.035 ± 0.009	30.552 ± 1.543
	EAPNet-MSRE	0.019 ± 0.005	33.174 ± 1.554
sAM ablation	EPNet-Bloss	0.029 ± 0.007	30.217 ± 1.479
	EAPNet-Bloss (sigmoid)	0.028 ± 0.006	30.275 ± 1.392

widening or deepening the contracting–expanding structure of UNet benefits in improved performance, it also results in a significant increase in the model size by 300% and 303%, respectively. Our EAPNet achieves the best performance with only a negligible increase in parameters and computational load, demonstrating its better suitability for addressing the optical inverse problem.

4.6. Attention module analysis

As presented Table 7 and Table S7, the network's performance degrades whether removing the entire sAM or just the softmax function, thereby fully confirming the effectiveness of the proposed sAM.

To better understand the benefits of our sAM, we visualized the attention maps of two samples from the RS dataset in Fig. 8, which indicate attention weights assigned to every pixel. It turns out that the sAM activated by the commonly used sigmoid function generates attention maps lacking discernible features and displaying a nearly uniform distribution Fig. 8(b). In contrast, our sAM, benefiting from the attention competition mechanism, yields attention maps possessing increased distinction Fig. 8(e). Moreover, we investigated the influence of loss functions on our sAM. MSE guides the sAM to assign more attention weights to inclusions with high absorption and their vicinities. When implemented with MSRE, our sAM produces larger attention weights in background areas. In contrast, the proposed Bloss assists our sAM in gaining broader and more uniform attention on inclusions, including deep-located tissue #5, despite its considerably weaker signal in the input p_0 images.

5. Discussion

We have conducted simulation validations, including multiple potential application scenarios. Besides achieving improved image-wise quantitative metrics, the proposed method delivers multifaceted advantages. Firstly, our method demonstrates impressive robustness to the intrinsic properties of targets (e.g., size, location, and relative absorption intensity), achieving superior consistency in tissue-wise accuracy. This suggests that our method has broader applicability and higher reliability in practical scenarios. Consistent accuracy in the spatial domain also enhances edge sharpness and target distinction, which is crucial for precise segmentation of regions of interest (ROIs) from the predicted image, thereby aiding further tasks like spectral unmixing. Further, the mean values within ROIs can be utilized as a statistically more reliable estimate if ROIs can be assumed optically homogeneous. The proposed Bloss plays a crucial role in the mentioned accuracy consistency. It effectively balances the contribution of each tissue to the loss function, enabling unbiased optimization of the entire image domain during training.

On the other hand, the EAPNet is a high-efficiency architecture for qPAT, which effectively enhances the performance baseline while maintaining a comparable level of network complexity to the conventional UNet (Table S8). The advance arises from two key components: the proposed E-P structure and sAM. In the E-P framework, spatial information capture and aggregation occur exclusively during the extraction phase, wherein the extractor learns global dependencies and generates contextual representation vectors for each pixel. Based on the extraction phase, the predictor can operate on a per-pixel basis and enhance nonlinear modeling capability in a parameter-efficient and computation-friendly manner. Consequently, the extractor and predictor primarily focus on global dependency and high nonlinearity, respectively, which suggests that the network, to some degree, deals with the two intractable features of the optical inverse problem in a stepwise manner, consequently reducing overall complexity. In addition, our sAM employs a novel attention competition mechanism, which aids the extractor in discerning valuable information from input images and its effective utilization in the predictor. Further, guided by the Bloss, our sAM can effectively attend to subtle yet crucial pixels in p_0 images (Fig. 8).

The phantom and in vivo experiments demonstrate the promising feasibility of our method in real-world scenarios. However, further in vivo research remains challenging. First, PA images are input data for models, while they often severely distort in practice, due to the non-ideal imaging system, as depicted in Fig. 7(a). Consequently, networks are forced to deal with PA image restoration, which is out of its intended task, resulting in degraded performance. Besides, low-quality regions, where noise overwhelms the signal, can contaminate the contextual representation vectors captured by the extractor and induce large prediction errors. To obtain accurate results, the quality of PA images should be carefully considered. Image restoration techniques may be applied beforehand if needed. Additionally, due to the lack of reliable in vivo measuring techniques for absorption coefficients, all experimental data of this study was processed by networks trained on simulated data. The simulated data is generated based on approximate models of physical processes, detector characteristics, and noise, and inevitably suffers a domain gap to real data. Consequently, networks cannot learn real-world information adequately. Recently, there has been rapid progress in generative networks and unsupervised learning, which may provide new opportunities for tackling this challenge [34, 45].

Another potential limitation is related to using the softmax as the activation function. In our sAM, the constant total attention is distributed to all pixels, including ones outside the imaging medium. These meaningless pixels may get a considerable portion of the attention weight when they contain significant noise levels. To mitigate it, a straightforward approach is to filter out signals originating from outside the imaging medium before inputting them into the network, as depicted in Fig. 7(a).

6. Conclusion

In this study, we propose a deep learning-based method for recovering the absorption coefficient distribution from PA images. The EAPNet proves to be a more suitable and efficient architecture than UNet, as it boosts the accuracy of predicted μ_a images with nearly identical network parameters and computational complexity. In addition, the Bloss effectively mitigates the difference in prediction accuracy between diverse tissues and consequently improves the applicability and reliability of our method in practice. Simulation and phantom experiment results demonstrate the improved performance of our method concerning both quantitative metrics and image quality. Our method also shows a promising result in a preliminary in vivo experiment. Our future work focuses on constructing high-quality experimental datasets to enable more complex real-world validations and thus facilitate the clinical translation process.

CRediT authorship contribution statement

Zeqi Wang: Writing – original draft, Software, Methodology, Conceptualization. **Wei Tao:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Hui Zhao:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was funded by National Natural Science Foundation of China, grant number 61873168.

Code availability

The code is provided in https://github.com/WZQ7/EAPNet, and the data involved in the main text is also provided.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.pacs.2024.100609.

References

[1] L.V. Wang, Multiscale photoacoustic microscopy and computed tomography, Nat. Photon. 3 (9) (2009) 503–509, http://dx.doi.org/10.1038/nphoton.2009.157.

[2] L.V. Wang, S. Hu, Photoacoustic tomography: In vivo imaging from organelles to organs, Science 335 (6075) (2012) 1458–1462, http://dx.doi.org/10.1126/science.1216210.

[3] M. Xu, L.V. Wang, Photoacoustic imaging in biomedicine, Rev. Sci. Instrum. 77 (4) (2006) 041101, http://dx.doi.org/10.1063/1.2195024.

[4] L.V. Wang, H.-i. Wu, Biomedical Optics: Principles and Imaging, John Wiley & Sons, 2012.

[5] T.D. Le, S.-Y. Kwon, C. Lee, Segmentation and quantitative analysis of photoacoustic imaging: A review, Photonics 9 (3) (2022) 176, http://dx.doi.org/10.3390/photonics9030176.

[6] M. Li, Y. Tang, J. Yao, Photoacoustic tomography of blood oxygenation: A mini review, Photoacoustics 10 (2018) 65–73, http://dx.doi.org/10.1016/j.pacs.2018.05.001.

[7] A. Taruttis, V. Ntziachristos, Advances in real-time multispectral optoacoustic imaging and its applications, Nat. Photon. 9 (4) (2015) 219–227, http://dx.doi.org/10.1038/nphoton.2015.29.

[8] A. Rosenthal, D. Razansky, V. Ntziachristos, Fast semi-analytical model-based acoustic inversion for quantitative optoacoustic tomography, IEEE Trans. Med. Imaging 29 (6) (2010) 1275–1285, http://dx.doi.org/10.1109/TMI.2010.2044584.

[9] B.E. Treeby, E.Z. Zhang, B.T. Cox, Photoacoustic tomography in absorbing acoustic media using time reversal, Inverse Problems 26 (11) (2010) 115003, http://dx.doi.org/10.1088/0266-5611/26/11/115003.

[10] M. Xu, L.V. Wang, Universal back-projection algorithm for photoacoustic computed tomography, Phys. Rev. E 71 (1) (2005) 016706, http://dx.doi.org/10.1103/PhysRevE.71.016706.

[11] J. Laufer, D. Delpy, C. Elwell, P. Beard, Quantitative spatially resolved measurement of tissue chromophore concentrations using photoacoustic spectroscopy: Application to the measurement of blood oxygenation and haemoglobin concentration, Phys. Med. Biol. 52 (1) (2007) 141–168, http://dx.doi.org/10.1088/0031-9155/52/1/010.

[12] B.T. Cox, J.G. Laufer, P.C. Beard, S.R. Arridge, Quantitative spectroscopic photoacoustic imaging: A review, J. Biomed. Opt. 17 (6) (2012) 061202, http://dx.doi.org/10.1117/1.JBO.17.6.061202.

[13] L. An, B.T. Cox, Estimating relative chromophore concentrations from multiwavelength photoacoustic images using independent component analysis, J. Biomed. Opt. 23 (07) (2018) 1, http://dx.doi.org/10.1117/1.JBO.23.7.076007.

[14] J. Gröhl, T. Kirchner, T.J. Adler, L. Hacker, N. Holzwarth, A. Hernández-Aguilera, M.A. Herrera, E. Santos, S.E. Bohndiek, L. Maier-Hein, Learned spectral decoloring enables photoacoustic oximetry, Sci. Rep. 11 (1) (2021) 6565, http://dx.doi.org/10.1038/s41598-021-83405-8.

[15] R. Hochuli, L. An, P.C. Beard, B.T. Cox, Estimating blood oxygenation from photoacoustic images: Can a simple linear spectroscopic inversion ever work? J. Biomed. Opt. 24 (12) (2019) 1, http://dx.doi.org/10.1117/1.JBO.24.12.121914.

[16] B.T. Cox, S.R. Arridge, K.P. Köstli, P.C. Beard, Two-dimensional quantitative photoacoustic image reconstruction of absorption distributions in scattering media by use of a simple iterative method, Appl. Opt. 45 (8) (2006) 1866, http://dx.doi.org/10.1364/AO.45.001866.

[17] S. Zhang, J. Liu, Z. Liang, J. Ge, Y. Feng, W. Chen, L. Qi, Pixel-wise reconstruction of tissue absorption coefficients in photoacoustic tomography using a non-segmentation iterative method, Photoacoustics 28 (2022) 100390, http://dx.doi.org/10.1016/j.pacs.2022.100390.

[18] B.T. Cox, S.R. Arridge, P.C. Beard, Gradient-based quantitative photoacoustic image reconstruction for molecular imaging, in: A.A. Oraevsky, L.V. Wang (Eds.), Biomedical Optics (BiOS) 2007, San Jose, CA, 2007, p. 64371T, http://dx.doi.org/10.1117/12.700031.

[19] R. Hochuli, S. Powell, S. Arridge, B. Cox, Quantitative photoacoustic tomography using forward and adjoint Monte Carlo models of radiance, J. Biomed. Opt. 21 (12) (2016) 126004, http://dx.doi.org/10.1117/1.JBO.21.12.126004.

[20] A.A. Leino, T. Lunttila, M. Mozumder, A. Pulkkinen, T. Tarvainen, Perturbation Monte Carlo method for quantitative photoacoustic tomography, IEEE Trans. Med. Imaging 39 (10) (2020) 2985–2995, http://dx.doi.org/10.1109/TMI.2020.2983129.

[21] A. Pulkkinen, B.T. Cox, S.R. Arridge, J.P. Kaipio, T. Tarvainen, A Bayesian approach to spectral quantitative photoacoustic tomography, Inverse Problems 30 (6) (2014) 065012, http://dx.doi.org/10.1088/0266-5611/30/6/065012.

[22] S. Mahmoodkalayeh, M. Zarei, M.A. Ansari, K. Kratkiewicz, M. Ranjbaran, R. Manwar, K. Avanaki, Improving vascular imaging with co-planar mutually guided photoacoustic and diffuse optical tomography: A simulation study, Biomed. Opt. Express 11 (8) (2020) 4333, http://dx.doi.org/10.1364/BOE.385017.

[23] A. Hussain, W. Petersen, J. Staley, E. Hondebrink, W. Steenbergen, Quantitative blood oxygen saturation imaging using combined photoacoustics and acousto-optics, Opt. Lett. 41 (8) (2016) 1720, http://dx.doi.org/10.1364/OL.41.001720.

[24] S. Guan, A.A. Khan, S. Sikdar, P.V. Chitnis, Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal, IEEE J. Biomed. Health Inf. 24 (2) (2020) 568–576, http://dx.doi.org/10.1109/JBHI.2019.2912935.

[25] H. Lan, D. Jiang, C. Yang, F. Gao, F. Gao, Y-Net: Hybrid deep learning image reconstruction for photoacoustic tomography in vivo, Photoacoustics 20 (2020) 100197.

[26] M. Guo, H. Lan, C. Yang, J. Liu, F. Gao, AS-Net: fast photoacoustic reconstruction with multi-feature fusion from sparse data, IEEE Trans. Comput. Imaging 8 (2022) 215–223.

[27] P. Chen, C. Liu, T. Feng, Y. Li, D. Ta, Improved photoacoustic imaging of numerical bone model based on attention block U-net deep learning network, Appl. Sci. 10 (22) (2020) 8089.

[28] L. Chao, P. Zhang, Y. Wang, Z. Wang, W. Xu, Q. Li, Dual-domain attention-guided convolutional neural network for low-dose cone-beam computed tomography reconstruction, Knowl.-Based Syst. 251 (2022) 109295, http://dx.doi.org/10.1016/j.knosys.2022.109295.

[29] C. Cai, K. Deng, C. Ma, J. Luo, End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging, Opt. Lett. 43 (12) (2018) 2752, http://dx.doi.org/10.1364/OL.43.002752.

[30] G.P. Luke, K. Hoffer-Hawlik, A.C. Van Namen, R. Shang, O-Net: a convolutional neural network for quantitative photoacoustic image segmentation and oximetry, 2019, arXiv preprint arXiv:1911.01935.

[31] C. Bench, A. Hauptmann, B. Cox, Toward accurate quantitative photoacoustic imaging: Learning vascular blood oxygen saturation in three dimensions, J. Biomed. Opt. 25 (08) (2020) http://dx.doi.org/10.1117/1.JBO.25.8.085003.

[32] J. Li, C. Wang, T. Chen, T. Lu, S. Li, B. Sun, F. Gao, V. Ntziachristos, Deep learning-based quantitative optoacoustic tomography of deep tissues in the absence of labeled experimental data, Optica 9 (1) (2022) 32–41.

[33] Y. Zou, E. Amidi, H. Luo, Q. Zhu, Ultrasound-enhanced Unet model for quantitative photoacoustic tomography of Ovarian Lesions, Photoacoustics 28 (2022) 100420, http://dx.doi.org/10.1016/j.pacs.2022.100420.

[34] J. Gröhl, M. Schellenberg, K. Dreher, L. Maier-Hein, Deep learning for biomedical photoacoustic imaging: A review, Photoacoustics 22 (2021) 100241, http://dx.doi.org/10.1016/j.pacs.2021.100241.

[35] K.-T. Hsu, S. Guan, P.V. Chitnis, Comparing deep learning frameworks for photoacoustic tomography image reconstruction, Photoacoustics 23 (2021) 100271.

[36] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Vol. 9351, Springer International Publishing, Cham, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[37] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Vol. 11211, Springer International Publishing, Cham, 2018, pp. 3–19, http://dx.doi.org/10.1007/978-3-030-01234-2_1.

[38] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning where to look for the Pancreas, 2018, arXiv:1804.03999.

[39] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[41] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB, IEEE, 2020, pp. 1–7.

[42] T. Tarvainen, M. Vauhkonen, V. Kolehmainen, J.P. Kaipio, Finite element model for the coupled radiative transfer equation and diffusion approximation, Internat. J. Numer. Methods Engrg. 65 (3) (2006) 383–405, http://dx.doi.org/10.1002/nme.1451.

[43] F.M. Brochu, J. Brunker, J. Joseph, M.R. Tomaszewski, S. Morscher, S.E. Bohndiek, Towards quantitative evaluation of tissue absorption coefficients using light fluence correction in optoacoustic tomography, IEEE Trans. Med. Imaging 36 (1) (2017) 322–331, http://dx.doi.org/10.1109/TMI.2016.2607199.

[44] D. Piao, S. Patel, Simple empirical Master–Slave Dual-Source configuration within the diffusion approximation enhances modeling of spatially resolved diffuse reflectance at short-path and with low scattering from a semi-infinite homogeneous medium, Appl. Opt. 56 (5) (2017) 1447, http://dx.doi.org/10.1364/AO.56.001447.

[45] Z. Wang, W. Tao, H. Zhao, The optical inverse problem in quantitative photoacoustic tomography: A review, Photonics 10 (5) (2023) 487, http://dx.doi.org/10.3390/photonics10050487.

[46] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention u-net for lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 683–687.

[47] Q. Fang, D.A. Boas, Monte Carlo simulation of photon migration in 3D turbid media accelerated by graphics processing units, Opt. Express 17 (22) (2009) 20178, http://dx.doi.org/10.1364/OE.17.020178.

[48] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, J. Ye, FIVES: A fundus image dataset for artificial intelligence based vessel segmentation, Sci. Data 9 (1) (2022) 475, http://dx.doi.org/10.1038/s41597-022-01564-3.

[49] T. Lyu, C. Yang, J. Zhang, S. Guo, F. Gao, F. Gao, Photoacoustic digital skin: Generation and simulation of human skin vascular for quantitative image analysis, 2022, arXiv:2011.04652.

[50] B.E. Treeby, B.T. Cox, K-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields, J. Biomed. Opt. 15 (2) (2010) 021314.

[51] R. Cubeddu, A. Pifferi, P. Taroni, A. Torricelli, G. Valentini, A solid tissue phantom for photon migration studies, Phys. Med. Biol. 42 (10) (1997) 1971–1979, http://dx.doi.org/10.1088/0031-9155/42/10/011.

[52] A. Buehler, A. Rosenthal, T. Jetzfellner, A. Dima, D. Razansky, V. Ntziachristos, Model-based optoacoustic inversions with incomplete projection data, Med. Phys. 38 (3) (2011) 1694–1704, http://dx.doi.org/10.1118/1.3556916.

[53] S.A. Prahl, Everything I Think You Should Know About Inverse Adding-Doubling, vol. 1, Oregon Medical Laser Center, St. Vincent Hospital, 2011, pp. 1–74.

[54] S. Zheng, H. Yingsa, S. Meichen, M. Qi, Quantitative photoacoustic tomography with light fluence compensation based on radiance Monte Carlo model, Phys. Med. Biol. 68 (6) (2023) 065009, http://dx.doi.org/10.1088/1361-6560/acbe90.

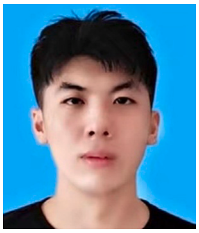
[55] A. Madasamy, V. Gujrati, V. Ntziachristos, J. Prakash, Deep learning methods hold promise for light fluence compensation in three-dimensional optoacoustic imaging, J. Biomed. Opt. 27 (10) (2022) http://dx.doi.org/10.1117/1.JBO.27.10.106004.

[56] A. Rodriguez-Molares, O.M.H. Rindal, J. D'hooge, S.-E. Måsøy, A. Austeng, M.A.L. Bell, H. Torp, The generalized contrast-to-noise ratio: A formal definition for lesion detectability, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (4) (2019) 745–759.

[57] K.M. Kempski, M.T. Graham, M.R. Gubbi, T. Palmer, M.A.L. Bell, Application of the generalized contrast-to-noise ratio to assess photoacoustic image quality, Biomedical Optics Express 11 (7) (2020) 3684–3698.

[58] D. Chen, J. Tachella, M.E. Davies, Equivariant imaging: Learning beyond the range space, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4379–4388.

[59] H. Lan, L. Huang, L. Nie, J. Luo, Cross-domain unsupervised reconstruction with equivariance for photoacoustic computed tomography, 2023, arXiv preprint arXiv:2301.06681.



Zeqi Wang received the B.E. degree in measurement and control technology and instruments from Sichuan University, Chengdu, China, in 2016.

He is currently working towards the Ph.D. degree in instrumentation science and technology with the School of Electronic Information and Electric Engineering, Shanghai Jiao Tong University, Shanghai, China.

Wei Tao received the B.S., M.S. and Ph.D. degrees in Instrument science and technology from Harbin Institute of Technology, Harbin, China, in 1997,1999 and 2003, respectively. She became a Professor of Shanghai Jiao Tong University in 2018. She is the author of three books, more than 100 articles and more than 40 inventions. Her research interests include opto-electronic measurement technology and application, methods and algorithms in vision measurement process, and laser sensors and measurement instruments.

Hui Zhao received the Ph.D. degree in the department of instrument engineering from Harbin Institute of Technology, Harbin, China, in 1996. Since 2000, he has been a Professor with the Department of Instrument Science and Engineering, Shanghai Jiao Tong University. His research interests include novel sensors and vision measurement method. He is the author of three books, more than 150 articles, and more than 50 inventions. He serves as the Vice Chair of Precision Mechanism Federation of China Instrument and Control Society, and the Vice Chair of Mechanical Quantity Measurement Instrument Federation of China Instrument and Control Society. He is a reviewer of IEEE Transaction on Instrument & Measurement, Sensors, Measurement, IEEE Sensors Journal, Optik, and several other journals.